A HEALTH ECONOMETRIC STUDY OF DIARY STUDIES OF AIR POLLUTION AND HEALTH

FINAL REPORT

BY

DAVID WYPIJ, PH.D.

JAMES H. WARE, PH.D.

JOEL SCHWARTZ, PH.D.

HARVARD SCHOOL OF PUBLIC HEALTH
677 HUNTINGTON AVENUE
BOSTON, MA 02115

NOVEMBER 1, 1988

Support for this research was provided by a Cooperative Agreement with the Environmental Protection Agency. The views expressed are those of the authors and not necessarily those of the Environmental Protection Agency.

EXTENDED SUMMARY

Diary studies, defined broadly as studies which record the health status of each study participant repeatedly over time, provide a powerful method for assessing the impact of short-term changes in the environment on human health. If health status is reported as the presence or absence of each of several symptoms, the data consist of sets of sequences of binary outcomes, one for each symptom and participant. The basic analytic objective, to estimate the exposure-response model linking exposure and symptom status, is complicated by the dependencies among responses on successive days (autocorrelation) and among responses of the same subject on different days (heterogeneity). This paper develops methods for analyzing diary data that address these complications and illustrates their use by analyzing data collected in a diary study conducted as part of the Six Cities Study of Air Pollution and Health.

The methods described in this report model the incidence rather than the prevalence of symptoms, with incidence defined as a positive report of symptom occurrence by an individual who did not report that symptom on the previous day. This strategy evolved from preliminary analyses which established that the respiratory symptoms under study had low prevalence rates and which also suggested that the incidence and recurrence of symptoms might have different mechanisms requiring different explanatory models.

When the endpoint is symptom incidence, only subjects free of the symptom on the previous day are at risk. This fact suggested a relatively simple analysis, in which response rates on successive days were treated as independent observations. Analyses using this approach are called ordinary logistic regression in this report.

Further investigation of the residuals from ordinary logistic regression established, however, that the incidence rates had detectable autocorrelation. Section 3 describes methods for modeling autocorrelation in the incidence data. The methods are applied to the Six Cities data in Section 4.

Similarly, it is reasonable to expect variability among subjects in the frequency of symptoms and, possibly, in sensitivity to air pollution exposures. Methods for modeling individual heterogeneity are also described in Section 3. Two-stage methods, which begin by modeling the data for each study participant and then analyze the parameters from these analyses, are not indicated when incidence rates are low as in the present application. Because they can be very useful in studies with higher background rates, however, they are also discussed in Section 3.

Finally, a comprehensive model would include both individual effects and autocorrelation. Models with these characteristics and the associated estimation procedures are described in Section 3.

Section 5 describes two strategies for the analysis of diary studies, one appropriate for data sets where each subject has relatively few events, and one appropriate to the setting where the data for each subject can be modeled separately.

1. INTRODUCTION

The estimation of the economic benefits of reductions in air pollutant concentrations requires quantitative information about the health effects of exposure to air pollutants. One possible pattern of health effects is increases in respiratory illnesses and symptoms during or shortly after periods of increased air pollution. Diary studies, in which participants report their respiratory health status repeatedly over time, are a natural method for studying the respiratory health effects of short-term fluctuations in air quality. Although diary studies are easy to design, they are difficult to analyze. This project was devoted to the development and empirical assessment of new statistical methods for analyzing diary data.

We consider the situation in which participants report the presence or absence of a respiratory symptom on each day. Thus, the data consist of many sequences of binary responses, each associated with a common sequence of environmental data. Because participant characteristics may determine sensitivity to environmental exposure, the methods described consider these characteristics in the analysis.

A diary study conducted as part of the Six Cities Study of Air Pollution and Health is described in Section 2. Section 3 describes several different families of models and associated statistical methods for the analysis of diary data. In Section 4, we apply these methods to a subset of the diary study collected in the Six

Cities Study. In Section 5, we discuss the strengths and weaknesses of the different methods, as well as guidelines for choosing a method to use in other settings.

2. DESCRIPTION OF THE DATA SET

The Six Cities Study of Air Pollution and Health is a large longitudinal study of the effects of exposure to air pollutants on the respiratory health of both children and adults (Ferris et al 1979, Ware et al 1984). The initial cohorts of children and adults were enrolled in six U.S. cities (Watertown, MA, Kingston-Harriman, TN, St. Louis, MO, Portage, WI, Steubenville, OH, and Topeka, KA) between 1974 and 1979. A second cohort of approximately 6,000 third to fifth grade school children was enrolled between 1982 and 1986. A subset of approximately 1,800 children from the second cohort was enrolled in a year-long diary study in which parents completed a daily report on the child's respiratory (and other) symptoms. For logistical reasons, the diary study extended over four school years (1984-1988). While the diary study was underway in one of the cities, air pollution concentrations were measured on a daily or more frequent basis. The ten symptoms monitored by the parents are listed in Appendix 1 along with a sample of the diary form and instructions to the parents. Parental smoking and gas stove information, as well as the child's respiratory illness history, were obtained via questionnaire.

Two decisions made early in the analysis played an important role. First, a literature review and discussion with specialists in respiratory disease led to the decision to use two or more respiratory symptoms to define three respiratory symptom complexes. These complexes, labeled upper respiratory infection (URI), lower respiratory inflammation (LRI), and rhinitis, are defined in Appendix 1. We also studied the frequency of cough *without other symptoms*. The ten symptoms (including any report of coughing) and the four constructed variables made a total of fourteen endpoints to be studied. Second, we distinguished between the *incidence* of a symptom, where incidence was defined as a report of the symptom by a child who *had not* reported that symptom on the previous day, and the *recurrence* of a symptom, where recurrence was defined as a report by a child who *had* reported that symptom on the previous day. This strategy was adopted because both medical considerations and preliminary data analysis suggested that the incidence and recurrence of symptoms might have different mechanisms requiring different explanatory models.

The incidence rates for all of the symptoms and symptom complexes were low, ranging from 0.2% (URI) to 1% (cough with or without symptoms). This implied that the data on recurrence of symptoms on days subsequent to a first report were very sparse. Thus, this report focuses on the analysis of incidence rates, where the incidence rate on a given day is the rate of reporting of a symptom among children who were free of that symptom on the previous day. Other more restrictive definitions of incidence were considered, but had little effect on the analysis because of the consistently low rate of symptom reporting.

3. MODELS FOR THE ANALYSIS OF DIARY DATA

This section describes methods for analyzing sequences of incidence rates when the objective is to model the effects of temperature, air pollution, and other time-varying variables on the incidence rate. Mismodeling the mean or the covariance structure of the sequences can lead to misleading results about environmental risk. Because data sets may be large and models complex, we seek computational simplicity, generally based on likelihood maximization or least squares methods.

The data consist of sequences $\{(\mathbf{x}_j, Y_j),\ 1 \leq j \leq T\}$, where $\mathbf{x}_j' = (x_{j1},\ldots,x_{jp})$ is a vector of p covariates affecting all subjects in the study at the j<u>th</u> occasion and Y_j is the number of incident cases of the symptom at the j<u>th</u> occasion among the n_j subjects who were symptom-free at the previous occasion. Y_j is assumed to have a binomial distribution with parameters n_j and p_j, where p_j is the marginal probability of symptom incidence for a subject.

This discussion focuses on the logistic model and its extensions. The logistic model is often used to model binary or binomial outcomes because the parameters can be interpreted as the logarithms of odds ratios and because computing is relatively simple. The logistic model is defined by

$$p_j = \exp[\boldsymbol{\beta}'\mathbf{x}_j] \ / \ (1 + \exp[\boldsymbol{\beta}'\mathbf{x}_j]),$$

or equivalently,

$$\text{logit}(p_j) = \beta' x_j.$$

Both the number of subjects at risk at any occasion, n_j, and the total number of occasions, T, can be large in diary studies.

The goal of the analysis is to estimate the effects of the pollution variables on incidence rates while controlling for other factors, particularly autocorrelation and subject heterogeneity. Autocorrelation (or serial correlation) is a particular form of the tendency for incidence rates close together in time to be positively correlated. This could be due to state dependence across individuals (e.g., symptoms may occur because other subjects had the symptom on the same or previous days), and/or time-dependent omitted covariates (which tend to be highly correlated in time).

Heterogeneity, or variability among individuals in the probability of response, induces positive correlation among responses on the same individual. It can be due to observable or unobservable within-subject covariates (such as smoking level or illness history) which vary across individuals, or different thresholds, susceptibilities, or reporting behavior across individuals. The latter could occur, for example, if participants varied in the severity of symptoms considered reportable.

Many methods have been proposed for modeling heterogeneity in binary data. One widely-used approach, the individual intercept model, assumes that the response probability follows a parametric model which depends on a linear function of covariates, and that one of those covariates is an intercept that varies from subject to subject. These intercepts may be treated either as fixed or random effects. This is equivalent to assuming that the response probability at some standard value of the covariates varies among subjects. This approach is adopted in this report, which focuses on logistic regression models with variable intercepts.

Failure to account for either autocorrelation or heterogeneity in the analysis can lead to errors in inference similar to those resulting from the naive use of standard methods in problems involving misspecified covariates, missing data, or measurement errors. In particular, mismodeling can result in failure to detect important effects as a consequence of biased point and interval estimates and incorrect hypothesis testing. Diary data typically have positively correlated outcomes, yielding less information than the same number of independent responses, so at a minimum the usual standard error estimates may need to be inflated.

3.1. Modeling Autocorrelation

It is natural to begin the analysis of incidence rates with models that assume independence of symptom rates on different days. Preliminary analysis of data from the Six Cities Study established, however, that residuals from regression models including important covariates were autocorrelated, and that this autocorrelation could not be explained by other measured time-varying covariates. Thus, refinements of the model were needed to account for this autocorrelation. This section describes several methods for modeling autocorrelation.

3.1.1. Using Lagged Prevalence or Incidence to Adjust for State Dependence

One possibility is that the probability of symptom occurrence on a given day depends on the participants symptom status on previous days. When modeling the prevalence of symptoms, Muenz and Rubinstein (1985), Cox (1970), and Korn and Whittemore (1979) used symptom status on the previous day as a covariate. This approach is not relevant to the analysis of incidence data, however, because all subjects at risk were, by definition, symptom free on the previous day. As an extension of this idea, however, one could assume that the probability of symptom occurrence for a study participant depends on the symptom status of others in the population on previous days. This dependence could arise if, for example, the symptoms were due to infectious diseases and risk of infection increased with the prevalence of the disease. Such epidemic or clustering effects could be modeled by assuming that

$$p_j = \exp[\beta' x_j + \rho x_{j,p+1}]/(1 + \exp[\beta' x_j + \rho x_{j,p+1}])$$

where $x_{j,p+1}$, the added covariate (with index p + 1) is the lagged prevalence or incidence rate in the study population. To use the lagged incidence rate as a covariate, we set $x_{j,p+1} = Y_{j-1}/n_{j-1}$. Alternatively, one could use the lagged prevalence rate in the study population as a measure of the likelihood of exposure to an infectious disease on a previous day.

The technique of including lagged prevalence or incidence rates in the model should be used cautiously, especially when assessing the weak effect of an autocorrelated environmental variable. The pollutant variable under study may also be autocorrelated, and the resulting collinearity will cause bias toward 0 in the coefficient of the pollutant variable if lagged prevalence is added to the model. Adding lagged prevalence or incidence to the model is only justified if there is a biological rationale for doing so, as with certain infectious diseases.

3.1.2. Using Residuals to Modify the Response Probabilities

Observed autocorrelation in incidence rates need not be due to state dependence. Suppose, for example, that there is a time-dependent omitted covariate. In general, such time-dependent variables have an autocorrelation structure of their own which induces autocorrelation in the residuals of the incidence model. As the residuals not only include a random component but are also a function of the omitted variable, the residuals (or a function of the residuals) can serve as a surrogate for the omitted covariate (Box and Jenkins 1970).

If the n_j are relatively large, using the central limit theorem we have that approximately

$$Y_j/n_j \sim N(p_j,\ p_j(1-p_j)/n_j).$$

If the errors, $((Y_j/n_j) - p_j)$, are autocorrelated, modifying the marginal probabilities based on an autoregressive model may be appropriate. As before, let

$$p_j = \exp[\boldsymbol{\beta}'x_j]/(1 + \exp[\boldsymbol{\beta}'x_j])$$

and let

$$p_j^* = p_j + \rho((Y_{j-1}/n_{j-1}) - p_{j-1})$$

define the conditional probability of incidence given the incidence rate on the previous occasion. Because of the heteroscedasticity of symptom rates, the model should be altered slightly to give

$$p_j^* = p_j + \rho(\sigma_j/\sigma_{j-1})((Y_{j-1}/n_{j-1}) - p_{j-1})$$

where $\sigma_j^2 = p_j(1-p_j)/n_j$. Preliminary work suggests that this modification avoids the need for restrictions on the admissible range of values of ρ, but this issue is still under investigation. These models can be generalized to include second or higher order autoregressive terms.

Another possibility is to assume an additive contribution on the logit scale. In particular, we could model

$$\text{logit}(p_j^*) = \boldsymbol{\beta}'\mathbf{x}_j + \rho(\sigma_j/\sigma_{j-1})((Y_{j-1}/n_{j-1}) - p_{j-1})$$

This model clearly imposes no restrictions on the allowable range of ρ. Further obvious modifications could be made to accommodate heteroscedasticity of the residuals and/or higher order autoregressive terms.

Still a third possibility is to extend the methods of empirical logits in a way that corresponds to the treatment of autocorrelated errors in the linear model. If we define the empirical logit as

$$Z_j = \ln[(Y_j + 0.5)/(n_j - Y_j + 0.5)]$$

we have that

$$Z_j \sim N(\boldsymbol{\beta}'\mathbf{x}_j,\ V_j)$$

approximately, where $V_j = (n_j+1)(n_j+2)/n_j(Y_j+1)(n_j-Y_j+1)$ (see Cox 1970). Considering the residuals on the logit scale $(Z_j - \boldsymbol{\beta}'\mathbf{x}_j)$ it is reasonable to consider the model

$$\text{logit}(p_j^*) = \boldsymbol{\beta}'\mathbf{x}_j + \rho(\sigma_j/\sigma_{j-1})(Z_{j-1} - \boldsymbol{\beta}'\mathbf{x}_{j-1})$$

with obvious extensions to accommodate higher order autoregressive terms. Again there are no restrictions on the allowable range of ρ. The use of empirical logits is recommended only when both the n_j are relatively large and the response rates are not too extreme. The empirical logits approach was not considered in the analysis of the Six Cities data because symptom rates were consistently low and zeroes were common.

In practice, it can be difficult to determine which autoregressive scheme is "best". The choice may sometimes be influenced by the availability of statistical software. The choice may influence parameter interpretation. The $\boldsymbol{\beta}$ parameters have more of a marginal interpretation when the residual effects are added on the probability scale and more of a conditional interpretation if the effects are added on the logit scale. Each of these schemes has the desired effect of reducing autocorrelation of the residuals.

3.1.3. Covariance Models to Accommodate Autocorrelation Effects

Most, if not all, of the methods described thus far for modeling autocorrelation lead to changes in the interpretation of regression coefficients for the variables under study, because these coefficients become partial regression coefficients adjusted not only for other covariates but for the residuals included in the model. Liang and Zeger (1986) and Zeger and Liang (1986) have described methods for fitting logistic models to the symptom rates while taking account of the correlation among symptom rates on different days. The general approach is to use a "working" correlation matrix to construct weighted point estimators and covariance estimators that are consistent and asymptotically normally distributed even when the covariance matrix is misspecified. Their methods are more efficient than estimators assuming independence of all the observations. The case of first or higher order autocorrelation represents a special case of their method which promises to be very useful in the analysis of diary data. Speaking more generally, these robust estimators and covariance matrix estimates deserve more serious attention in many epidemiological applications.

3.2 Models for Heterogeneity

The previous section described methods for modeling autocorrelation of incidence rates. This section focuses on the effects of individual heterogeneity, assuming that the residuals are not autocorrelated in time. In Section 3.3, we discuss models which combine the autocorrelation and subject heterogeneity effects.

The simplest method for analyzing incidence is to combine the responses from all subjects at each occasion. Observations from different individuals are usually assumed to be independent, so if there are n_j subjects at risk, then the number of affected subjects at the j<u>th</u> response time, Y_j, is assumed to follow a binomial distribution with parameters n_j and

$$p_j = \exp[\beta' x_j]/(1 + \exp[\beta' x_j]),$$

where $\mathbf{x}_j$ denotes the (common) covariate level at the j<u>th</u> response time. This greatly simplifies the computing, which depends in the collapsed data only on the order of T, the number of response times, not on the number of subjects. This method may be appropriate when a sample of homogeneous subjects is chosen randomly from a population of interest.

One may wish to focus on particular subgroups, perhaps defined by one or more categorical subject characteristics. In this case, a stratified analysis may be appropriate, with a separate parameter vector for each stratum. Data for subjects in the same stratum can still be collapsed, so the calculations remain manageable.

If the subject characteristics are continuously distributed or include several categorical covariates, more general methods are needed. We introduce additional notation to discuss these methods. We now focus on individuals, and consider sequences of the form $\{(\mathbf{x}_{ij}, Y_{ij}), 1 \leq i \leq N, 1 \leq j \leq T\}$, where $\mathbf{x}'_{ij} = (x_{ij1},\ldots,x_{ijp})$ is a known vector of p covariates and Y_{ij} is a binary random variable. We set $Y_{ij} = 1$ when the symptom of interest is incident in the i<u>th</u> subject at the j<u>th</u> occasion, and set $Y_{ij} = 0$ if the subject is at risk and the symptom is not observed. The covariate vectors, $\mathbf{x}_{ij}$, can involve both subject characteristics and time-varying covariates. Typically, the time-varying covariates, such as environmental factors, affect all subjects, so many of the components of $\mathbf{x}_{ij}$ will not vary over subjects. The marginal distribution of Y_{ij} is assumed to depend on $\mathbf{x}_{ij}$ and we write

$$p_{ij} = P[Y_{ij} = 1 \mid \mathbf{x}_{ij}]$$

for the marginal probability of symptom incidence on the j<u>th</u> occasion for the i<u>th</u> subject. The methods we describe allow for unequal numbers or timing of observations for each subject, but for simplicity we assume that all subjects are seen at exactly the same occasions.

3.2.1. Random Effects Models (Varying Slopes and Intercepts)

If T is large, we may observe heterogeneity among subjects in response rates that is not explained by the within-subject covariates. Subjects may also vary in their sensitivity to pollutant exposures, as measured by the regression coefficients. Korn and Whittemore (1979) proposed a two-stage analysis based on the assumption that each subject's sequence of binary responses follows a logistic model but with coefficients that vary among subjects. Specifically, they assume a parameter vector, $\boldsymbol{\beta}_i$, for individual i, so that the conditional probability of response for the i<u>th</u> subject at the j<u>th</u> response time is given by

$$p_{ij} | \boldsymbol{\beta}_i = \exp[\boldsymbol{\beta}_i' \mathbf{x}_{ij}]/(1 + \exp[\boldsymbol{\beta}_i' \mathbf{x}_{ij}]).$$

They then assume that the $\boldsymbol{\beta}_i$ arise from a multivariate normal distribution. Their estimation technique also proceeds in two stages. First, estimate $\hat{\boldsymbol{\beta}}_i$ for the i<u>th</u> subject using ordinary logistic regression. If the number of observations on the i<u>th</u> subject is sufficiently large, the asymptotic distribution of $\hat{\boldsymbol{\beta}}_i$ is approximated by

$$\hat{\boldsymbol{\beta}}_i | \boldsymbol{\beta}_i \sim N(\boldsymbol{\beta}_i, \hat{\mathbf{V}}_i),$$

where $\hat{\mathbf{V}}_i$ is the usual information-based variance-covariance matrix. Then, in the second stage, we assume

$$\boldsymbol{\beta}_i | \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

and so

$$\hat{\boldsymbol{\beta}}_i | \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma} + \hat{\mathbf{V}}_i),$$

and $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are then estimated from the averages and sums of squares of the $\hat{\boldsymbol{\beta}}_i$ and $\hat{\mathbf{V}}\text{i}$ using the method of moments.

Here $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are the population parameters and are viewed as the primary parameters of interest. The above method indirectly accounts for within-subject covariates through the variation in the coefficients. Weighted least squares regression could also be used to regress the individual regression coefficients, $\hat{\boldsymbol{\beta}}_i$, on characteristics of the children or their families, such as parental smoking or the child's illness history.

This two-stage estimation method is relatively easy to implement, but has two drawbacks. First, the asymptotic normality assumption of $\hat{\boldsymbol{\beta}}_i | \boldsymbol{\beta}_i \sim N(\boldsymbol{\beta}_i, \hat{\mathbf{V}}_i)$ holds only when there is a sufficiently large number of observations per subject and the response rate for each is sufficiently high. In other cases the model is suspect. In particular, they are not appropriate when response rates are very low, as is the case in the Six Cities Diary study. In fact, for consistency and asymptotic normality of the estimates we need that both T → ∞ and N → ∞. Second, this estimation method is not the most efficient. More efficient (but more computationally intensive) multivariate random effects models are available (Stiratelli, Laird, and Ware 1984).

3.2.2. Random Intercepts Models (Common Slopes and Varying Intercepts)

An alternative approach is to assume that the regression coefficients are constant across subjects but that each subject has a different underlying response rate (as measured by the intercept). This formulation allows individual heterogeneity due to observed or unobserved subject covariates, differences in reporting, or other reasons, but information regarding $\boldsymbol{\beta}$, the primary parameter vector of interest, is strengthened by combining information across subjects.

In particular, we postulate that responses from the i<u>th</u> subject follow the logistic model with success probability

$$p_{ij}|\alpha_i = \exp[\alpha_i + \boldsymbol{\beta}'\mathbf{x}_{ij}]/(1 + \exp[\alpha_i + \boldsymbol{\beta}'\mathbf{x}_{ij}])$$

where α_i denotes the intercept for the i<u>th</u> subject. There are basically three estimation approaches.

First, if one is not interested in the individual intercepts, a conditional maximum likelihood approach can be used. Here the likelihood contribution of the i<u>th</u> subject is given by the extended hypergeometric distribution

$$L_i(\beta) = \frac{\prod_{y_{ij}=1} \exp[\beta' x_{ij}]}{\sum_J [\prod_{j \varepsilon J} \exp[\beta' x_{ij}]]}$$

where J is the set of subsets of {1,....,T} with size equal to $\sum_j y_{ij}$. This likelihood also arises in the analysis of matched or stratified binary observations or 2 x 2 tables. The major virtue of maximizing the conditional likelihood

$$L(\beta) = \prod_{i=1}^{N} L_i(\beta)$$

is that this estimator is consistent and asymptotically normal in both the large strata and sparse strata cases, i.e., whenever $T \to \infty$ or $N \to \infty$ (or both). The major difficulty is that programs to compute conditional maximum likelihood estimates do not accept the immense amount of data arising from diary studies. For example, if a subject is followed daily for one year and has twenty days of symptom incidence, then there are $\binom{365}{20} \cong 4.26 \times 10^{32}$ terms to be summed in the denominator of this subject's contribution to the likelihood.

Second, if one can make the assumption that the α_i, $1 \leq i \leq N$, are i.i.d. observations from a distribution $f(\alpha_i | \tau)$, one can assume a mixture model for the random effects. Here the likelihood contribution for the ith subject is an integral of the form

$$L_i(\boldsymbol{\beta},\tau) = \int_{\alpha_i=-\infty}^{\infty} \left(\prod_{j=1}^{T} \frac{\exp[(\alpha_i + \boldsymbol{\beta}'x_{ij})\cdot y_{ij}]}{1 + \exp[\alpha_i + \boldsymbol{\beta}'x_{ij}]} \right) f(\alpha_i|\tau)\,d\alpha_i$$

One common choice of $f(\alpha_i|\tau)$ is the univariate normal distribution.

Because the integration cannot be performed analytically, this method would require numerical integration or some other approximation. The estimates will be consistent and asymptotically normal as $N \to \infty$. A minor problem here, as for the conditional maximum likelihood estimator, is that the individual intercepts, α_i, are not estimated. However, in either case, once the fixed effects, $\boldsymbol{\beta}$, are estimated a second stage analysis could be used to estimate the α_i using likelihood or empirical Bayes techniques.

Finally, an approach that has not been discussed much in the literature is to use (unconditional) maximum likelihood to jointly estimate the parameters $\boldsymbol{\beta},\alpha_1,\ldots,\alpha_N$. Here the likelihood contribution of the i<u>th</u> individual is given by

$$L_i(\boldsymbol{\beta},\alpha_i) = \prod_{j=1}^{T} \frac{\exp[(\alpha_i + \boldsymbol{\beta}'x_{ij}) \cdot y_{ij}]}{1 + \exp[\alpha_i + \boldsymbol{\beta}'x_{ij}]}$$

and the full likelihood $L(\boldsymbol{\beta},\alpha_1,\ldots,\alpha_N) = \prod_{i=1}^{N} L_i(\boldsymbol{\beta},\alpha_i)$ is maximized over the N+p parameters. Standard logistic regression packages require iterations involving the inversion of an (N+p) x (N+p) matrix, which is too difficult for diary studies when N is large. However, due to the special structure of the data, an efficient "two-step" maximization scheme can be implemented. The idea is as follows: first, given fixed values for $\alpha_1,...,\alpha_N$, determine $\boldsymbol{\beta}$ to maximize the full likelihood $L(\boldsymbol{\beta},\alpha_1,...,\alpha_N)$. Then, given a fixed value for $\boldsymbol{\beta}$, determine α_i to maximize $L_i(\boldsymbol{\beta},\alpha_i)$ for each individual. Repeat these steps until convergence. This algorithm is easy to implement, and convergence will not be a problem due to the concavity of the logistic regression likelihood. The variance-covariance matrix can be estimated in the usual way as long as the block structure of the problem is used to simplify the required matrix inversion.

The computations involved with this method are easier than those for the conditional maximum likelihood or random intercept approaches. In addition, the estimated intercepts (and their variance estimates) can be used in a second stage analysis using weighted least squares to assess the effects of subject covariates (such as smoking levels or illness history). For consistency and asymptotic normality of our estimates we need that $T \rightarrow \infty$. This is not too restrictive an assumption for diary studies in which subjects are followed for a long period of time (such as daily for one year or longer). Empirical Bayes modifications (Duffy and Santner 1987, Wypij 1988) may improve inference with this method.

3.3. Models which Accommodate Both Autocorrelation and Subject Heterogeneity Effects

The analyses summarized in the next section indicate that both autocorrelation and heterogeneity are present in the Six Cities diary data. This seems likely to be true in other studies as well. Thus, models which include both heterogeneity and autocorrelation are needed. The most promising approach appears to be that which combines fixed intercepts which vary over subjects with the Liang and Zeger technique for modeling autocorrelation. Work is underway to assess the feasibility of this approach, and if appropriate, to develop the software necessary to implement this method.

4. RESULTS

This section illustrates some of the methods discussed in the previous section by applying them to diary data from the Six Cities Study. The analyses are based on data from three cities, Watertown, MA, Kingston-Harriman, TN, and St. Louis, MO. Data for the other three cities are still being collected and processed. In Section 4.1, we document the presence of autocorrelation in the daily incidence rates and discuss methods for controlling for this autocorrelation in the analysis. In Section 4.2, we discuss heterogeneity of response and how it can be treated in the analysis. Finally, in Section 4.3, we explore the data for lagged effects of pollutant concentrations on symptom rates.

4.1 Evidence for Autocorrelation of Incidence Rates

Autocorrelation of time series data should be considered only after controlling for the effects of measured covariates. In the Six Cities diary data, only one independent variable, temperature, had strong and consistent effects on symptom rates. In all analyses discussed in this report, the effects of temperature were controlled by introducing temperature and the square of temperature into the regression model. Each pollutant was investigated separately while controlling for the effects of temperature in this way. Analyses not reported here established that seasonal variables did not contribute significantly to the model after the two temperature terms had been added. Here we consider one set of analyses, those investigating the effects of sulfur oxide concentration on the incidence of cough (with or without symptoms) in Watertown.

Figure 1 shows the partial autocorrelation function of the daily incidence rates. Figure 2 shows the partial autocorrelation function of residuals from an ordinary logistic regression model for cough incidence including temperature, temperature squared, and sulfur dioxide concentration. (The partial autocorrelation of order k is the correlation between $y(t)$ and $y(t-k)$ after controlling for $y(t-1), \ . \ . \ . \ . \ y(t-k+l)$.) Autocorrelation is reduced by inclusion of the explanatory variables, but there is a strong indication of, at a minimum, first and second order autocorrelation in the residuals. The autocorrelation may be due to unmeasured time-dependent covariates. Epidemic effects, which can be represented as lagged values of prevalence, may also be important. The second panel in Figure 2 shows the partial autocorrelation function after fitting a regression model with first-order autoregressive errors to the data. (Using the multiplicative AR model. See below.) This plot suggests the presence of second order autocorrelation.

As noted in the methods section, this autocorrelation can be modeled in several ways. Three approaches considered here are the additive AR model, which assumes that the differences between the observed and expected incidence rates have an autoregressive error structure, the multiplicative AR model, which assumes that the conditional probability of symptom incidence on a given day depends on the observed rates on previous days through an additive contribution to the linear model for the logit of symptom probability, and the Liang-Zeger model, which assumes that the marginal distributions of symptom incidence follow a multiple logistic distribution, while the errors have an autoregressive covariance structure. We have considered each of these possibilities, while also considering the possibility that the symptom probabilities depend on the previous day's disease prevalence.

Table 1 shows the effect on the regression coefficient for sulfur dioxide concentration of choosing several different models for the error structure. Sulfur dioxide was a significant predictor of cough incidence in an ordinary logistic regression model assuming independent errors. Models that adjusted for autoregressive errors tended to reduce the statistical significance of the SO_2 coefficient. In the multiplicative AR model, a first-order autoregressive term was significant but lagged prevalence was not, suggesting that the autoregressive model satisfactorily explains the dependency of the incidence rate on the previous day's outcomes (data not shown). In the Liang and Zeger and additive AR models, the first-order term was not significant. (See Table 1.) Since these models are slightly different, it is not surprising that they give different results for the order of the autoregression.

Perhaps the most important feature of Table 1 is the consistency among the estimated regression coefficients and standard errors for sulfur dioxide obtained by different methods. Even though the autocorrelation among successive days was of moderate size and highly significant, different approaches to modeling this autocorrelation, including ignoring it entirely (as ordinarily logistic regression does), had little effect on the results.

4.2 Individual Effects

One potentially attractive way to account for individual variability is to perform separate regressions on each subject. We call this the Korn and Whittemore (KW) approach. Despite the reservations described in the methods section, we examined this approach for our data. KW also allowed us to examine the relations between individual intercepts and child-specific covariates, such as presence of chronic respiratory disease and parental smoking. In Watertown, the weighted mean of the individual SO_2 coefficients for cough incidence was close to the coefficient obtained from the analysis of the daily incidence rates. In Kingston and St. Louis, the KW approach showed a stronger association than the grouped analysis. This raises the possibility that methods allowing individual variation to weak effects are more sensitive than methods for analyzing pooled data. Nevertheless, the low incidence rates made KW inappropriate for these data. The individual regressions failed to converge for about 10% of the subjects. In the remaining subjects, the distributions of the coefficients were clearly non-normal. Even after adjusting for the different weights assigned to different coefficients, a highly skewed distribution remained. Therefore, estimation and testing procedures based on normality assumptions do not apply.

KW did allow a preliminary investigation of child-specific risk factors. No association was found between parental smoking and the underlying rate of cough incidence, but doctor-diagnosed asthma, lung function less than 70% of predicted, and moisture in the basement did show associations. Because of the poor distributional properties of the estimated coefficients, we do not plan to continue with this approach. The suggestion of greater sensitivity to small changes due to air pollution indicates that consideration of individual variation may be important. The individual intercept model seems most capable of achieving this goal.

4.3 Lag Effects

Any effect of pollution exposure on symptoms is not necessarily contemporaneous. The lag between exposure and symptom may also differ among the pollutants, whose modes of action vary. For instance, Dockery **et al** (1982) reported a lag of 1-2 weeks between exposure to high levels of particulates and reductions in lung function. In contrast, Spektor **et al** (1988) and Kinney **et al** (1988) reported that high ozone exposure causes almost immediate reductions in lung function. To explore the lag relationship in the diary data, we utilized simple logistic regression with no autoregressive components. Temperature was modeled with a linear and quadratic term, as suggested by exploratory plots and analyses. The concurrent and lagged pollutant measures for up to 14 days lag were examined individually. If the pattern in these individual regressions suggested a model using a weighted linear combination of pollutant concentrations on several previous days, such a distributed lag model was also fit. This approach was applied in each of three cities (Watertown, Kingston-Harriman, and St. Louis).

These analyses showed the strongest associations between Upper Respiratory Illness and acid measurements (Table 2). The acid measurements on the two previous days had the largest regression coefficients. A model using a weighted combination of concentrations on the three previous days had the largest coefficient, about twice as large as that for any single day.

5. STRATEGIES FOR THE ANALYSIS OF DIARY STUDIES

The analyses described in this report have shown that rates of incidence of symptoms among participants in a diary study tend to be autocorrelated, perhaps because of epidemic effects and the effects of omitted covariates on response rates. Moreover, our work and the work of others has shown that subjects have heterogeneous response rates. Analyses of diary data should examine the effects of both autocorrelation and heterogeneity on estimated regression coefficients and their standard errors.

This report has described several methods for investigating heterogeneity. These methods focus on the daily incidence rates of the symptom under study and use one of several approaches to modeling the autocorrelation among these rates. In the example used to illustrate these methods, controlling for autocorrelation had little effect on the results. Additional work is needed to determine the conditions under which autocorrelation can be disregarded.

Methods for modeling heterogeneity of response rates were outlined in Section 3.2. These methods have yet to be applied to the diary data from the Six Cities Study. Future work will focus on evaluating the computational feasibility of the three estimation methods described in Section 3.2. The simultaneous treatment of autocorrelation and heterogeneity requires an extension of these methods which is still under development.

An alternative approach, the two-stage method (Korn and Whittemore 1979), can be used to model both autocorrelation and heterogeneity when each individual has both a substantial number of reporting days and a substantial number of days with symptoms. In this method, one begins with modified logistic regression of the data for each subject, using the quasi-likelihood method of Liang and Zeger (1986) with a working covariance matrix that represents autocorrelation among the residuals. At the second stage, one analyzes the individual regression coefficients by generalized least squares methods, as described in Section 3. This method is efficient, relatively easy to implement, and easy to explain. It is not appropriate, however, for endpoints with low incidence like those investigated in the Six Cities Study.

REFERENCES

Box, G.E.P., and Jenkins, G.M. (1970). Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco.

Cox, D.R. (1970). The Analysis of Binary Data. Metheun, London.

Dockery, D.W., Ware, J.H., Cook, N.R., Speizer, F.E., Herman, S. and Ferris, B.G. Jr. (1982). Change in Pulmonary Function in Children Associated with Air Pollution Episodes. J. Air Pollution Control Assoc. 32, 937-942.

Duffy, D.E., and Santner, T.J. (1987). Estimating Logistic Regression Probabilities. Statistical Decision Theory and Related Topics, Volume IV, 31-51.

Ferris, B.G., Speizer, F.E., Spengler, J.D., Dockery, D., Bishop, Y.M.M., Wolfson, M., and Humble, C. (1979). Effects of Sulfur Oxides and Respirable Particles on Human Health: Methodology and Demography of Populations in Study. Am. Rev. Resp. Dis. 120, 767-779.

Kinney, P.L., Ware, J.H., Spengler, J.D., Dockery, D.W., Speizer, F.E., and Ferris, B.G. Jr. (1988). Short-term Pulmonary Function Change in Association with Ozone Levels. To appear in Am. Rev. Resp. Dis.

Korn, E.L., and Whittemore, A.S. (1979). Methods for Analyzing Panel Studies of Acute Health Effects of Air Pollution. Biometrics 35, 795-802.

Liang, K.-L., and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika 73, 13-22.

Muenz, L.R., and Rubinstein, L.V. (1985). Markov Models for Covariate Dependence of Binary Sequences. Biometrics 41, 91-101.

Spektor, D.M., Lippman, M., Lioy, P.J., Thurston, G.D., Citak, K., James, D.J., Bock, N., Speizer, F.E., and Hayes, C. (1988). Effects of Ambient Ozone on Respiratory Function in Active, Normal Children. Am. Rev. Resp. Dis. 137, 313-320.

Stiratelli, R., Laird, N., and Ware, J.H. (1984). Random-Effects Models for Serial Observations with Binary Response. Biometrics 40, 961-971.

Ware, J.H., Spiro, A., Dockery, D.W., Speizer, F.E., and Ferris, B.G. Jr. (1984). Passive Smoking, Gas Cooking, and Respiratory Health of Children Living in Six Cities. Am. Rev. Resp. Dis. 129, 366-374.

Wypij, D. (1988). Pseudotable Methods for the Analysis of 2 x 2 Tables. Submitted for Publication.

Zeger, S.L., and Liang, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. Biometrics 42, 121-130.

Table 1

Regression coefficients for cough incidence on SO_2 in Watertown, MA for different model specifications

Model	$\beta(SO_2)$	S.E.	Comments
Ordinary logistic	0.0133	0.0052	
<u>Models with Two Autoregressive Terms</u>			
Multiplicative AR	0.0116	0.0053	
Liang and Zeger	0.0132	0.0059	AR(1) insignificant
Additive AR	0.0117	0.0053	AR(1) insignificant
Lagged prevalence (No AR(1) term)	0.010	0.0056	Lagged prevalence insignificant
<u>Reduced Models</u>			
Liang and Zeger (AR(2) only)	0.0130	0.0059	
Additive AR (AR(2) only)	0.0113	0.0052	

Table 2

Evidence for Lagged Effects of H_2SO_4 Concentrations on Upper Respiratory Symptoms

	Coefficients				
Lag Period	0	1	2	3	Distributed Lag
WAT	0.431	0.683	0.690	0.076	1.28
KH	0.121	0.461	0.258	0.232	1.16
STL	0.171	0.848	0.524	0.227	2.34

Figure Legends

Figure 1. Sample autocorrelation function of the daily incidence rates of cough.

Figure 2. The left panel shows the sample partial autocorrelation function of the residuals from an ordinary logistic regression of cough incidence rates on temperature, the square of temperature, and sulfur dioxide concentrations. The right panel shows the autocorrelation function when the logistic regression function is modified to assume that the errors have first-order autoregressive error structure.

FIGURE 1

PARTIAL AUTOCORRELATION FUNCTION

LOGIT OF COUGH INCIDENCE

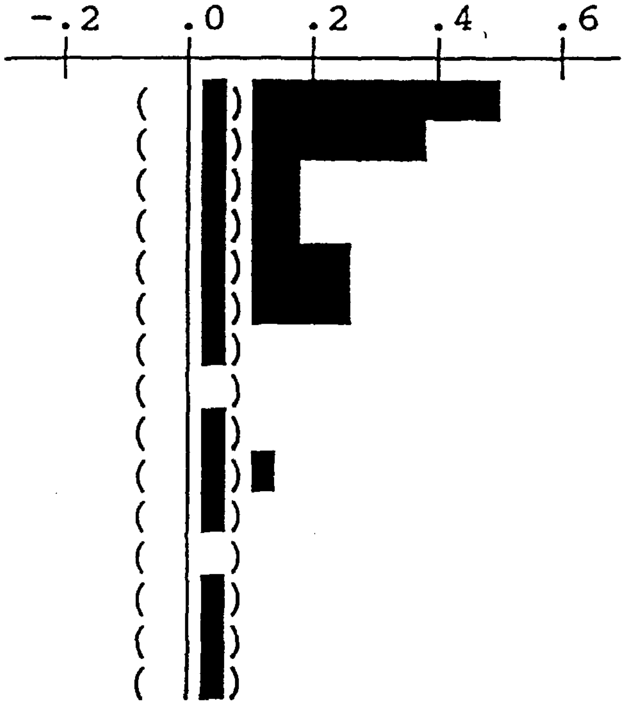
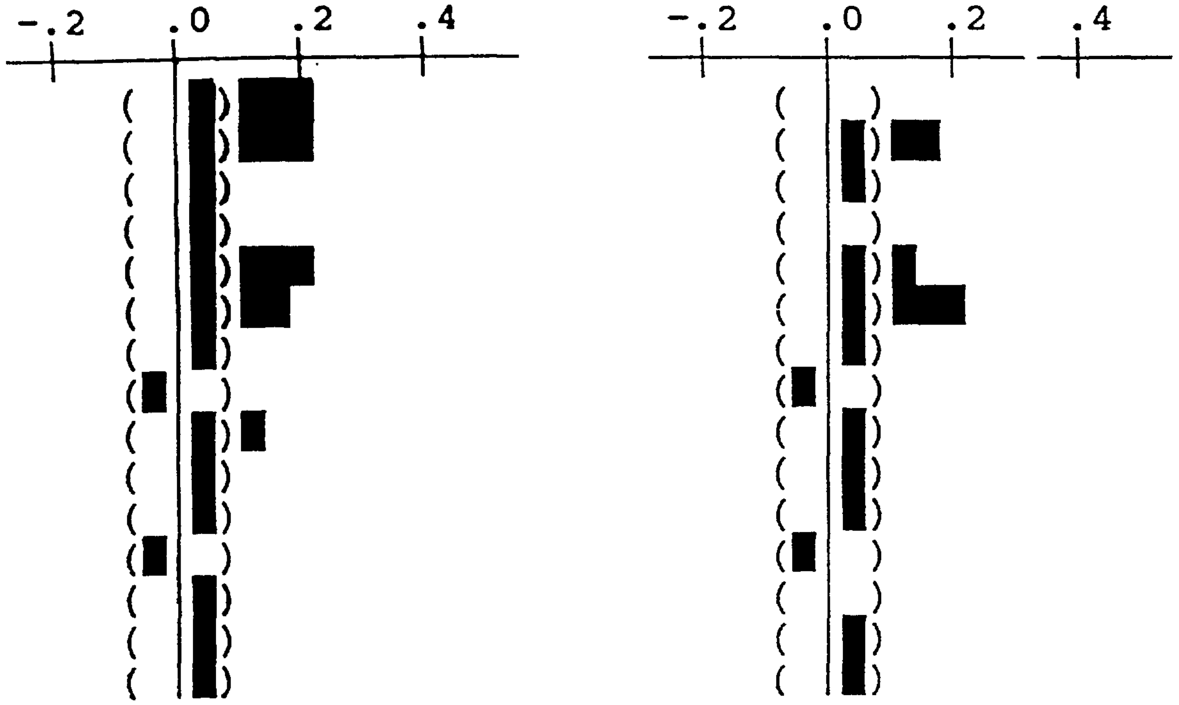


FIGURE 2



CONTROLLING FOR TEMPERATURE, TEMPERATURE SQUARED, AND SO2

ADDING AN AR1 TERM

APPENDIX 1

Instructions for Using Calendars

Attach your calendar to a bulletin board or kitchen cabinet. On any day that your child has any of the symptoms listed on the top of each page, simply write the code letter(s) corresponding to his or her symptom(s) under that day's date. Be sure to also write in the number code (1, 2 or 3) or asterisk (*) whenever they might apply. Please review the sample calendar for August on the next page; it is included here to give you an idea of what your health calendar might look like once it is filled out.

The Guide to Health Code Letters (see last page) will help you become familiar with the health code we are using in the study. Read through this Guide right away, and keep it handy as a reference aid whenever you might have any questions on how to code a particular symptom or sick day.

Every two weeks we shall phone you and ask you to read the calendar record to us. At the end of each month simply tear off the page for the completed month and mail it back to us in a prepaid envelope we shall send you. We want you to keep this daily health diary of your child through several changes of season, throughout the school year and summer of 1987.

Harvard Health/Air Quality Study Group
Harvard School of Public Health
665 Huntington Avenue; Building #1, Room 1414
Boston, MA 02115

Telephone: (617) 732-0830 -- Call Collect

SAMPLE CALENDAR: AUGUST

If your child has any of the symptoms listed below, write the code letters for each symptom in the block for that day. Please also add the number code (1, 2 or 3) or asterisk (*) whenever applicable.

A	Hoarseness	G	Fever
B	Sore Throat	H	Ear Pain or Discharge
C	Cough	J	Runny or Stuffed Nose
D	Phlegm from the Chest	K	Burning, Aching or Red Eyes
E	Pain in the Chest	S	Upset Stomach
F	Wheezing	O	None of the Above or Healthy

- - - - - - - - - - - - - - -

1	*Stayed Home for the Day*	3	*Hospitalized*
2	*Saw Doctor or Nurse*	*	*Out of Town Over Night*

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
					1 O	2 O
3 O	4 O	5 J	6 J	7 CJ1	8 CJ2	9 C J
10 C	11 C	12 O	13 O	14 O	15 O	16 O
17 O	18 O	19 O	20 O 2	21 O	22 O*	23 O*
24 O	25 S	26 O	27 O	28 O	29 O	30 O
31 O						

August 1986

A Hoarseness -- Hoarse or dry throat; raspy voice; laryngitis.

B Sore Throat -- Any soreness or irritation of the throat; “strep throat;” tonsilitis.

C Cough -- Sporadic, intermittent or protracted coughing.

D Phlegm from the Chest -- Phlegm or mucus coughed up from the lungs or area of the throat below the voice box; congestion in the lungs.

E Pain in the Chest -- Aching, irritation or feeling of constriction in the lungs.

F Wheezing -- Wheezing or whistling sound from the chest with or without shortness of breath.

G Fever -- An above-average temperature recorded by thermometer at any point during the day or evening.

H Ear Pain or Discharge -- Ear ache or ear infection; discharge of fluid from the ears.

J Runny or Stuffed Nose -- Nasal or sinus congestion; post-nasal drip; phlegm or mucus from back of throat; sneezing; itching of the nasal passages.

K Burning, Aching or Red Eyes -- Sensations of burning, itching or aching in the eyes or eyelids; red or watery eyes.

S Upset Stomach -- Stomach pain, vomiting, acid indigestion, or diarrhea.

O None of the Above or Healthy -- No symptoms at all, or symptoms other than those listed above (e.g., headache, rash, etc.).

----

1 *Stayed Home for the Day* -- Interruption in the child’s daily routine resulting from any illness; e.g. child stays home from school or on a non-school day remains indoors.

2 *Saw Doctor or Nurse* -- Any appointment or visit with a health practitioner, whether regularly scheduled or not.

3 *Hospitalized* -- Admitted to a hospital or clinic as an in-patient for one night or more.

4 *Out **of** Town Over Night* -- Out of town five miles or more over night or longer, for vacation, holidays or any other reason.

Guide to Health Code Letters

DEFINITIONS OF SYMPTOM COMPLEXES USED IN THE ANALYSIS OF THE DIARY DATA

Upper Respiratory Illness: Any two of hoarseness, sore throat, and fever (symptoms A, B, and G, respectively).

Lower Respiratory Illness: Any two of cough, chest pain, phlegm, and wheeze (symptoms C, D, E, and F, respectively).

Rhinitis: Runny nose (symptom j), with no other symptom